Big Data: Philosophy, emergence, crowdledge, and science education

Renato P. dos Santos
renatopsantos@ulbra.edu.br

PPGECIM- Doctoral Program in Science and Mathematics Education,
ULBRA-Lutheran University of Brazil, Brazil

Abstract. Big Data already passed out of hype, is now a field that deserves serious academic investigation, and natural scientists should also become familiar with Analytics. On the other hand, there is little empirical evidence that any science taught in school is helping people to lead happier, more prosperous, or more politically well-informed lives. In this work, we seek support in the Philosophy and Constructionism literatures to discuss the realm of the concepts of Big Data and its philosophy, the notions of 'emergence' and *crowdledge*, and how we see *learning-with-Big-Data* as a promising new way to learn Science.

Keywords: Science education, philosophy of Big Data, learning-with-Big-Data, emergence, crowdledge

Introduction

The growing number of debates, discussions, and writings "demonstrates a pronounced lack of consensus about the definition, scope, and character of what falls within the purview of Big Data" (Ekbia et al., 2015). The usual 'definition' of Big Data as "datasets so large and complex that they become awkward to work with using on-hand database management tools" (Snijders, Matzat, & Reips, 2012) is unsatisfactory, if for no other reason than the circular problem of defining "big" with "large" (Floridi, 2012). In (dos Santos, 2016), we have already discussed a possible comprehensive definition of Big Data.

Among the many available definitions of Big Data, we find the following one a most insightful for the purposes of this study:

"*Big data is more than simply a matter of size; it is an opportunity to find insights in new and emerging [emphasis added] types of data and content, to make your business more agile, and to answer questions that were previously considered beyond your reach*" (IBM, 2011).

We do not foresee reasons for a decrease in the production of digital data and, therefore, are confident Science and Economics will continue to require Big Data Analytics.

This work begins with a discussion on the realm of the concepts of Big Data and its philosophy. In the sequence, the idea of 'emergence' is covered and we defend our thesis that Big Data Analytics is better understood by the concept of *crowdledge* than by the overused 5-V definition. At last, we will sum up by discussing how we see *learning-with-Big-Data* as a promising way to learning Science.

Big Data

Relying on Gartner's hype cycle for emerging technologies analysis (Gartner Group, 2015), which presents Big Data already having crossed the 'peak of inflated expectation' and being

now more than a mere marketing buzzword, Swan (2015) maintains that Big Data would already have passed out of hype and arrived as a field that deserves a Philosophy of its own.

However, as this paper focus on *learning-with-Big-Data* as a promising new way to learn Science, we will concentrate on its epistemological issues (knowledge) rather than ontological (existence and meaning) or axiological (valorisation, including ethics and morality) ones.

Having too much data is not the real, epistemological problem with Big Data and, therefore, its solution is not of more and better techniques and technologies to "shrink" it back to a manageable size (Floridi, 2012), as many vendors propose, but in developing a philosophical investigation of the methods of Big Data Analytics.

Traditionally, the analyst tested a theory by studying relevant data by means of a previously selected appropriate method based on her knowledge of the techniques and data; now, hundreds of different algorithms can be applied to a dataset to determine the best or a composite model or explanation, seeking to gain insights 'born from the data' (Kitchin, 2014). Big Data Analytics is thus considered to enable a radically different epistemological approach for making sense of the world (Kitchin, 2014).

This new epistemological approach led Gray, back on 2007, to argue that we are witnessing the emergence of what he named *eScience*, a "new, fourth paradigm for science" (Hey, Tansley & Tolle, 2009), succeeding and unifying the Empirical, the Theoretical, and the Computational ones (Gray, 2007) (Table). According to Gray (2007), it is based on the exploration and analysis, by means of data management and statistics, of data previously captured by instruments or generated by simulators and stored in databases.

However, as the word *paradigm*, since its introduction by Kuhn (1962), became "banal (not to mention *paradigm shift*!)" (Hacking, 2012, p. xi) and is nowadays "embarrassingly everywhere" (2012, p. xix), the proposal of Big Data being a 'fourth paradigm' for science is very questionable.

For Frické (2015), several examples of *data-driven science*, as Kitchin (2014) calls it, such as the Sloan Digital Sky Survey, The Ocean Observatories Initiative, as well as a good portion of biodiversity science or genomics science, are, methodologically speaking, merely gathering more data, not in themselves offering any explanations or theories, solving scientific problems, or aiming to do anything of that nature. For Frické (2015), data-driven science, as a 'fourth paradigm,' "is a chimera".

Table 1. The four paradigms of science (Source: Kitchin, 2014)

Paradigm	Nature	Form	When
First	Experimental Science	Empiricism; describing natural phenomena	pre-Renaissance
Second	Theoretical Science	Modelling and generalization	pre-computers
Third	Computational Science	Simulation of complex phenomena	pre-Big Data
Fourth	Exploratory Science	Data-intensive; statistical exploration and data mining	Now

On another interpretation, Big Data is not seen as a new paradigm, but rather as a return to a new era of Empiricism. In a provocative article in *Wired Magazine*, with the quite drastic title *The End of Theory: The Data Deluge Makes The Scientific Method Obsolete* (2008), Anderson went "for a reaction" (Norvig, 2009) and affirmed:

"*The scientific method is built around testable hypotheses. These models, for the most part, are systems visualized in the minds of scientists. The models are then tested, and experiments confirm or falsify theoretical models of how the world works. This is the way science has worked for hundreds of years, [but it] "is becoming obsolete""* (Anderson, 2008).

According to Anderson and Prensky, scientists no longer have to make educated guesses, construct hypotheses, coherent models, unified theories, or really any mechanistic explanation at all, and test them with data-based experiments and examples. Instead, they can throw complete, digitally stored data sets into the biggest computing clusters the world has ever seen without hypotheses about what it might show and let statistical algorithms find patterns and produce scientific conclusions without further experimentation where science cannot (Anderson, 2008; Prensky, 2009). According to Anderson (2008), petabytes of data allow one to say, "Correlation is enough" and supersedes causation.

First of all, we think, it would be appropriate to ask about *what* scientific method Anderson is talking. The definite article 'The' in the expression 'The Scientific Method' seems to imply there is one and only one, unique scientific method. On the contrary to the popular view, however, almost all modern historians and philosophers of the natural sciences would agree there is no such thing as a single, unique scientific method (Woodcock, 2014); the whole idea that data always trumps theory, that the process is mechanical, and that it is individualistically objective, is "a myth" (Bauer, 1994, p. 20). Instead, scientists themselves think of the plurality of methods available as a 'toolbox' (Wivagg & Allchin, 2002) for different purposes: scientific reasoning, mathematical methods, computational methods, measurement methods, and experimental methods (Woodcock, 2014). Consequently, whatever is 'the' method Anderson is thinking about, it is far from "becoming obsolete."

According to Ekbia et al., the introduction of Big Data seems to develop along a trajectory that repeats the same historical pattern of flouting "previously upheld criteria" (van Fraassen, 2008, p. 277) "that has been the hallmark of scientific change for centuries" (Ekbia et al., 2015).

The first "upheld" criterion is the *Common Cause Principle*, proposed by Hans Reichenbach (1956), that asserts that being A and B two simultaneous and positively correlated events, then one of the events causes the other or there is a further event C occurring before A and B which is a *common cause* of A and B. Boyd & Crawford (2012) warns that those enormous quantities of data offer connections that radiate in all directions, and it is too easy to incur in the practice of *apophenia*: seeing patterns where none actually exists. Without a theoretical explanation for the observed correlation between A and B, it is impossible to know whether this is really a pattern or a mere illusion caused by the external *common cause* C.

The second criterion threatened by the "correlation is enough" statement would be the *Appearance-from-Reality* criterion, proposed by van Fraassen (2004), that affirms that "*a complete Physics must explain how […] appearances are produced in reality*". However, data is typically gathered using measuring instruments and sensors, which are constructed or adopted in the light of the theories available, about what we are measuring and how the instruments work (Frické, 2015). Therefore, Big Data is ultimately *theory-laden* (Frické, 2015), and correlation does not seem to be 'enough'! Knowing the correlative results of data analytics, and the properties that a particular system has, is not sufficient. The science of data science must develop the predictive and prescriptive causal models that are the basis of science-driven understanding, at a systems level, of what causes those properties (Fox &

Hendler, 2014). As we will see later, the search for "what causes" the observed correlations is a central point in our proposal of *learning-with-Big-Data*.

Now, that same Gartner analysis also affirms Internet of Things (IoT) is rapidly approaching that peak and IoT will soon be of huge importance to business too (Gartner Group, 2015).

IoT can be understood as a linked set of computer programs that process data collected by a network of "things" widespread all over the world, which, contrary to people (O'Leary, 2013), do not incur in limitations of time, attention, and accuracy (Ashton, 2009). Those "things" are physical or virtual objects – from roadways to pacemakers –embedded with electronics, software, sensors, and connectivity to enable them to exchange data with the manufacturer, operator and/or other connected devices (Rathod et al., 2015). Besides, every piece of data collected by them has a geographic location and time stamp, as time and space correlations processing is an important part of IoT (Chen, Mao & Liu, 2014).

'Things' can even be aware of other 'things,' may communicate with other 'things', and can gather information and knowledge from their interactions with other 'things' (O'Leary, 2013). Furthermore, the 'Internet of Things' is beginning to be configured to include inputs from humans linked to the Internet, ultimately being referred to as the 'Internet of Everything' (O'Leary, 2013).

Contemplating this change, we agree with Swan that such a Philosophy of Big Data might be helpful in conceptualizing and realizing Big Data Science as a service practice, and in transitioning to data-rich futures (Swan, 2015a), with multiple human minds or entities (e.g.; artificial intelligence) co-existing in mutually growing, full-fledged productive and generative cloudmind collaborations coordinated via the Internet cloud (Swan, 2016).

In the next section, the concept of 'emergence', which is sometimes associated with the 'new' types of data and content provided by Big Data, will be discussed.

Emergence

Interest in the concept of 'emergence' has been renewed in our present days due to discussions of the behaviour of complex systems, mental causation, intentionality, and consciousness (O'Connor & Wong, 2015); this section provides an introductory account of it (For a concise review, see (O'Connor & Wong, 2015)).

The term 'emergence' comes from the Latin verb *emergo*, which means "to arise, to rise up, to come up, or to come forth" (Vintiadis, 2013). However, British emergentists of the late-nineteenth and early twentieth centuries, whose members mainly include J.S. Mill, C.D. Broad, and S. Alexander, were possibly the first to accord a philosophical status to this term since the English philosopher G.H. Lewes coined it on 1875:

"*the emergent is unlike its components insofar as these are incommensurable, and it cannot be reduced to their sum or their difference*" (Lewes, 1875, p. 412).

Emergent phenomena may be roughly conceptualized as the novel and coherent macro-level structures, patterns, and properties that arise during the process of self-organization in complex systems, in contrast to the micro-level, more fundamental components and processes out of which they arise; yet, they are neither predictable from, deducible from, nor reducible to the parts alone (Goldstein, 1999). It is sometimes said that consciousness is the only one clear case of an emergent phenomenon (Chalmers, 2006; O'Connor & Wong, 2015).

The concept of emergence is "a perennial philosophical problem" and is far from having a "single, specific, useful, pre-theoretical" definition (Bedau, 2002). It is a term used in different

ways both in the philosophy of mind and in the natural and cognitive sciences (Vintiadis, 2013). Trying to encompass this difference, Bedau (1997) introduced the notion of 'weak emergence', distinct from 'strong' emergence, most common in philosophical discussions of emergence (see (Bedau & Humphreys, 2008) and (Chalmers, 2006) for comprehensive studies on this concept).

Strong emergence occurs when macro phenomena are systematically determined by low-level facts without being deducible from those facts (Chalmers, 2006). In weak emergence, macro properties are ontologically and causally reducible in principle to the micro low-level properties of the parts of the system, even if they are new, unexpected, and different from these properties, and the reductive micro-explanation is extraordinarily complex (Bedau, 2008; Chalmers, 2006). The weak one is also supposed to be more frequent than the strong one (Bedau, 1997) and is considered vital to understanding all sorts of phenomena in nature and, in particular, biological, cognitive, and social phenomena (Chalmers, 2006).

Emergence is also subject to metaphysical objections (Bedau, 2002). For Goldstein (1999), it looks more like a merely descriptive term pointing to the patterns, structures, or properties exhibited on the macro-level than an explanation for them.

According to Crutchfield (1994), while contemporary physics does have the tools for detecting and measuring complete order and ideal randomness, there are no physical principles that define and dictate how to identify emergent patterns. Therefore, one needs to ask continually whether the 'newness' of the pattern is not merely a result of the ever-ready mechanism of projecting patterns onto the world, built into our epistemological, cognitive apparatus (Goldstein, 1999). For Goldstein (1999), it is this lack of sufficient frameworks for grasping emergent order that still interferes with accepting emergents as having an ontological status.

On the other hand, Wan (2009) argues that it is ultimately problematic to treat 'emergence' as an epistemological category as it would involve the "epistemic fallacy" (Bhaskar, 1975) in confusing our explanation or prediction of novel qualities with the novel qualities in question themselves. Accordingly, Bunge affirms the concept of emergence is ontological, not epistemological (Bunge & Mahner, 1997, p. 29), and warns that, contrary to a widespread opinion, emergence has nothing to do with the possibility or impossibility of explaining qualitative novelty (Bunge, 2003, p. 21). To which, he adds "Emergence is often intriguing but not mysterious: explained emergence is still emergence." After careful analysis, Wan (2009) demonstrates that epistemology-centred approaches to emergence are generally, in fundamental ways, "out of step with contemporary science", while an ontological construal of emergence, such as Bunge's one, can provide a more viable theoretical basis for scientific investigations.

Crowdledge

In our understanding, the central parameter that distinguishes Big Data from previous data analysis processes is not the overused 5-V definition (Beulke, 2011; Hurwitz et al., 2013; Laney, 2001), but *crowdledge*, which we define as the knowledge that [weakly] emerges – and is, therefore, unexpected - from Big Data (and IoT) analysis of individuals' digital footprints spontaneously left in the digital universe we live in by means of web searches, postings, 'shares' and 'likes' in social networks, phone calls, images and videos uploaded to sharing websites, etc., as well as through a myriad of 'things' such as pacemakers, radio-frequency identification (RFID) tags, wearables, IOT sensors, connected cars, and smarthome and smartcity sensors.

However, as 'things' evolve in gathering knowledge from their interactions with other 'things' and humans linked to the Internet (O'Leary, 2013), and non-humans' intelligence abilities develop (Swan, 2015b), the definition above may have to refer to multi-species intelligent individuals (Swan, 2015c) in a near future.

We emphasize, however, that crowdledge is not the same as the *wisdom of crowds*. Woolley et al. (2010) found evidence that human groups (crowds) as a whole perform a wide variety of tasks with a success rate (wisdom) above and beyond what can be explained by knowing the abilities of the individual group members. Its application in the business world and the conditions that are necessary for the crowd to be wise were discussed in detail by James Surowiecki in his book *The Wisdom of Crowds* (2005). While not new to the information age, this process has been pushed into the mainstream spotlight by social information sites such as *Wikipedia* and *Yahoo!Answers* and other web resources that rely on human opinion (Baase, 1996).

On the other hand, crowdledge is not *collective wisdom*, also called group wisdom and co-intelligence. The idea of wisdom suggests the intelligence of a collective extending nor just through space (including many people) but through time as well (including many generations, thus making room for both experience and memory – as they are transmitted, for example, in proverbs and sayings (Landemore & Elster, 2012, p. 7).

Finally, crowdledge is not *collective intelligence* as well. According to Pierre Lévy (1997, p. 61), collective intelligence is a negotiation and decision-making process within heterogeneous and dispersed communities. Unlike collective wisdom, some authors (Bloom, 1995; 2000) consider that collective intelligence is not exclusively human, and it may be also associated with animal and plant life, an idea Lévy sees at the opposite of the sense he uses in his book (Lévy, 1997, p. 16).

An example of crowdledge could be the patterns observed in population movements following the Haiti 2010 earthquake and cholera outbreak estimated from position data of subscriber identity module (SIM) cards, as observed by Bengtsson et al. (2011).

Another example could be the example discussed below of the unexpected knowledge that the searches for the term 'wireless hotspot' were more frequent during maxima in solar activity (Figure 1) that emerged from Big Data analysis of individuals' spontaneous digital footprints left in Google searches.

Learning-with-Big-Data in Science Education

It should not be necessary to discuss the worthwhileness of Science Education again.

When one turns to how Science education is being done, however, the discussion becomes harsher. The "triumphal progress" of science literacy research has produced little empirical evidence that any science taught in school, from Newton's laws to natural selection, does help people lead happier, more successful, or more politically savvy lives (Feinstein, 2011). On the contrary, the current scientific illiteracy should be evident when one considers, for example, our present climate change denials, anti-vaccination movements, recurring end-of-world prophecies, and school boards discussions about the inclusion of creationism in their science curriculum (Zimmerman & Croker, 2014).

One possibility for this failure could be a bias Papert identified in our schools mindset against ideas (such as the nature of science and scientific reasoning) in favour of mere skills and facts (e.g., denatured laboratory practices and 'physical laws'), an approach that students experience as "excruciatingly boring" (Papert, 2000). Specifically, he observed that Jean

Piaget's very tough idea that all learning takes place by discovery was translated into school curricula as "discovery learning" and lost its intellectually adventurous side in the form of being made to 'discover' what someone else (and someone you may not even like) wants you to discover (and already know!) (Papert, 2000).

To the *Instructionism*, so present in our current schools, that focus on the transfer of knowledge to students, being essentially irrelevant whether via book, teacher, or tutorial program (Papert & Harel Caperton, 1991, pp. 7–10), Papert explicitly opposes his *Constructionism* (1980) (for a review, see (Parmaxi & Zaphiris, 2014)), which comprises an approach that touches on the epistemological issues of the production of knowledge by students, the nature of knowledge and the nature of knowing (Papert & Harel Caperton, 1991, pp. 8–10). Specifically, Papert proposed the computers being used as learning tools rather than teaching tools (1980, p. 19).

Constructionism is often seen as close connected with LOGO, the programming language well known for its turtle graphics developed at the Massachusetts Institute of Technology in the late 1960s. Because of this connection, Constructivism has inherited much criticism Logo received from a growing body of research (Parmaxi & Zaphiris, 2014). Papert minimized this criticism as a "technocentric", "poor way to talk about Logo" (Papert, 1987) that saw 'computers' and of 'LOGO' as agents that acted directly on thinking and learning (Papert, 1987). Furthermore, the importance of programming languages as the main semantics of Constructionism faded away with the advent of new interface technologies that offer dynamic manipulation of representations (Healy & Kynigos, 2009) and Constructionism stays relevant (Kynigos, 2012).

Papert considered the concrete thinking as where the most important work is to be done, while the School errs in a perverse commitment to spending minimal time there and moving as quickly as possible to the abstract thinking (Papert, 1993, p. 143). According to Papert, the slower development of a particular concept could be attributed to the relative lack of appropriate materials that would make it simple and concrete (Papert, 1980, p. 7). For example, few students can explain what kind of a thing is a 'law of motion' and whether there are other laws of motion besides Newton's ones (1980, p. 124).

Moreover, while Piagetian *Constructivism* sees children as active builders of knowledge (Papert, 1999), Papert's Constructionism expands it by attaching particular importance to the role of concrete buildings *in the world* – as a support for those *in the head* – in the form of a 'product' of a more public sort, which can be shown, discussed, examined, probed, and admired, be it a sand castle, a computer program, a poem, or a theory of the universe (Papert, 1993, p. 142). Wilensky (1991, p. 198) affirms that concreteness is not a property of the object; rather, we come into engaged relationship with the knowledge needed for its construction 'in the world', and it is then especially likely "that we will make this knowledge concrete" (1991, p. 202).

According to Papert, his goal has always been the design of objects that children could "make theirs for themselves and in their own ways", "objects-to-think-with" (Papert, 1980, p. 11). Here, this author seems to anticipate Rosa's idea of students *to-think-with* and *to-learn-with* the computer (Rosa, 2008), which we advanced to *learning-with-Big-Data* (dos Santos, 2014).

Instead of a mere training in computational infrastructure or predictive analytics, or referring to *computerized instruction*, *educational data mining*, or *learning analytics* (Mayer-Schönberger & Cukier, 2014), *learning-with-Big-Data* is based on Papert's Constructionism and is seen as a promising way to build scientific knowledge, to learn to do science, to learn to think like a scientist (dos Santos, 2014).

Higginbotham (2011) believes that, similarly to what happened to broadband, computers, electricity, and other significant impacts in our culture, for Big Data really to become a force for change in our productivity, it will have to reach the masses. In fact, accessible software to ordinary people is already emerging, trying to channel large amounts of data in a more human understanding, such as *Tableau*, *Karmasphere*, *Revolution Analytics, Microsoft Power Map*, and *HDInsight*, without the need to learn *Hadoop*, *MapReduce*, and other that shall arise soon. However, those accessible tools are useless without Big Data sets, and even if such sets can be obtained from the Census Bureau or digital capture of social and economic activity, there are not many freely accessible Big Data sets relevant to natural science learning.

On the other hand, Google search engine stores about one hundred billion Web searches monthly, all identified by time and place of origin, to be later used by its highly profitable advertising programs, such as *DoubleClick*, *Google Analytics*, *Google AdWords,* and *Google AdSense*, from which comes 90% of Google Inc. revenue (Google Inc., 2015). Fortunately, this stored information was also made available through various public analytical tools released the last few years, such as *Google Trends* (available at http://www.google.com/trends/), *Google Correlate* (available at http://www.google.com/trends/correlate), *Google Zeitgeist* (closed 22 May 2007 and replaced by *Hot Trends,* a dynamic feature in *Google Trends)*, and *Google Insights for Search* (merged into *Google Trends* on Sep. 2012).

According to this view, we started investigating the feasibility of using those public and free Big Data applications as mediators, the computer in Science learning.

Among previous proposals of using Big Data in teaching and learning, the following ones must be mentioned: Baram-Tsabari & Segev propose the use of *Google Trends*, *Google Zeitgeist*, and *Google Insights for Search* for research and discussion on the public understanding of science and the distinction between science and pseudoscience in the classroom (2009a, 2009b, 2015; 2012). Bülbül (2009) and Yin et al. (2013) suggest determining and discussing trends in Physics and Education through keyword searches via *Google*, *Google Scholar*, and *Google Trends*.

Google Correlate allows users to sort through several years of Google search queries from around the world to get a graphical plotting showing the popularity of particular search terms over time. It takes an input, which can be an individual search term, a temporal or spatial data series uploaded by the user, or a sketch of a graph made with its 'Search by Drawing' feature. It then finds the set of individual search queries whose spatial or temporal pattern are most highly correlated (measured by Pearson correlation coefficient R^2) with the input (Mohebbi et al., 2011). Numerous academic works based on *Google Correlate* can be found in the scientific literature in several domains, including Public health, Economics, Sociology, and Meteorology, among others. We believe there is a potential for *Google Trends* and *Google Correlate* to find unexpected, and even unusual, correlations that may, however, serve as clues to interesting phenomena from the pedagogical and even scientific point of view.

As an example, we have input in *Google Correlate* data of the weekly variation in solar activity (NOAA/NESDIS/NCEI, 2013), to which the best correlated (R^2 = 0, 7523) was 'wireless hotspot' (**Figure 1**Figure 1) (dos Santos, 2014). In the graph produced by *Google Correlate* (Figure 2), this correlation is fairly apparent.

A possible causal relationship for the observed correlation would be that maxima in solar activity affect radio-communication conditions (Vitinskii, 1965), hindering the reach of hotspots and, consequently, users accustomed to using certain hotspots would be forced to searches on Google for new hotspots to connect (dos Santos, 2014). We are of course aware of the statisticians' warning "correlation does not imply causation" (Field, 2003, p. 10) and the students were lead to deepen their research from various other sources to confirm or refute

the hypothesis, an effort that can be extremely productive in terms of science learning (dos Santos, 2014). Their research results became then a 'product' that could be constructionistly shown, discussed, examined, probed, and admired (Papert, 1993, p. 142) by the rest of the class by means of presentations.

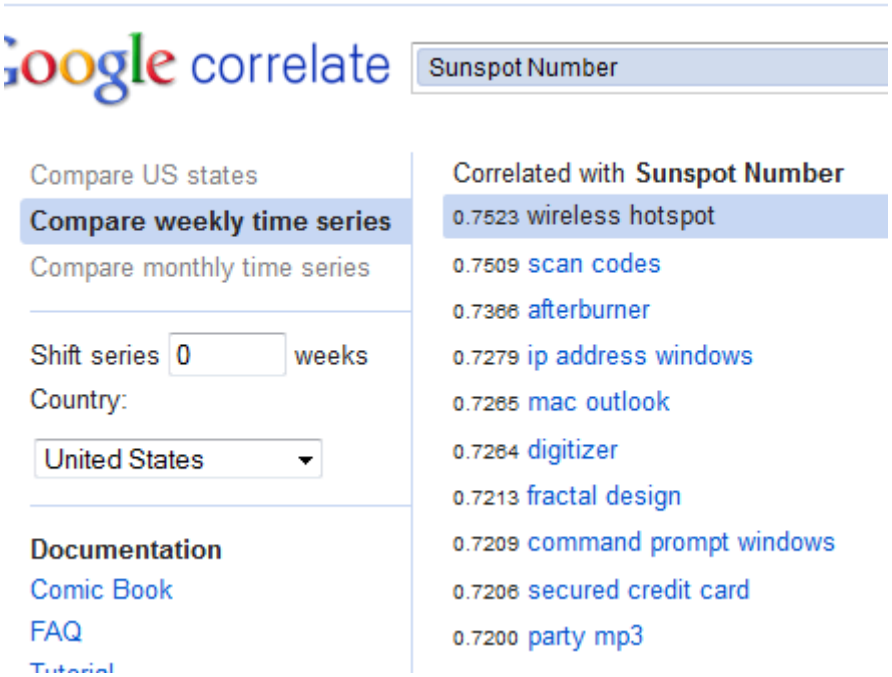

Figure 1. Search term frequencies on Google positively correlated with the weekly change in the sunspot number from January 5, 2003, to March 31, 2013. (Source: Google Correlate, http://www.google.com/trends/correlate).

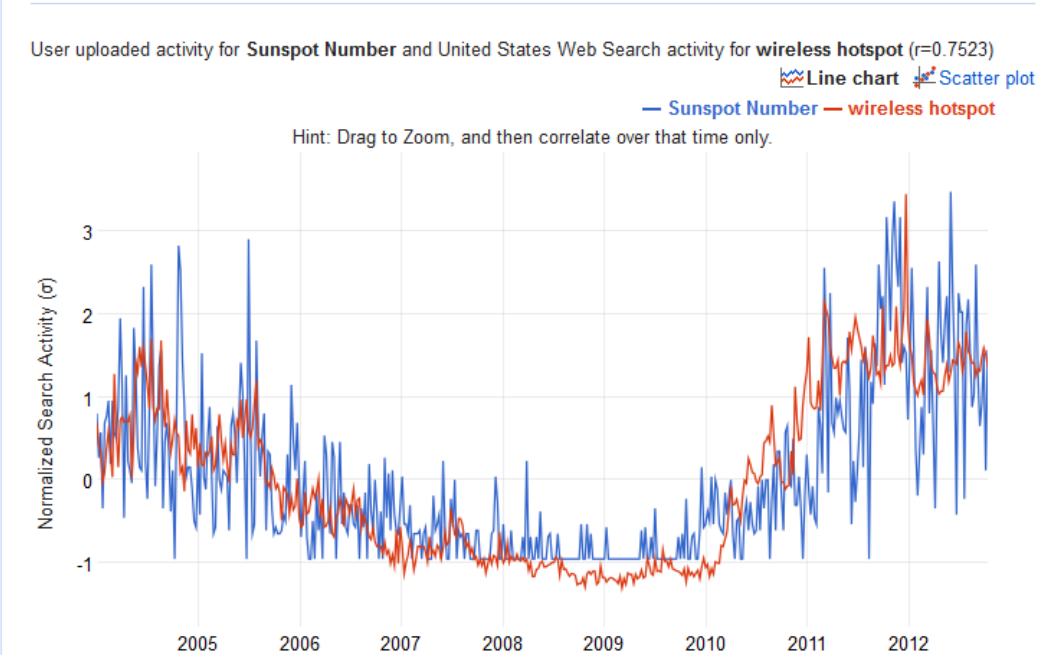

Figure 2. Comparison between the frequency of the search term 'wireless hotspot' in Google and the weekly change in the sunspot number from January 5, 2003, to March 31, 2013.

From an epistemological point of view, the student can be lead to see this correlation as a candidate to a 'physical law' of her own (even if of little scientific interest)! From a constructionist point of view, this is a rare chance of making such an abstract concept concrete. Here one sees the Big Data applications acting as mediators in Science Learning, a case of *learning-with-Big-Data*. Other examples are discussed in (dos Santos & Lemes, 2016).

In our understanding, this is also an example of *weakly emergence*, as discussed above: a new, unexpected, non-obvious property (the correlation between the weekly change in the sunspot number and the frequency of the search term 'wireless hotspot' in Google) of a complex system (Google users' seeking behaviour), distinct from the properties of the different parts of the system (individual Google users, unique Google searches) and caused by them; however, while being *a posteriori* deductible from those different parts of the system, it would be virtually impossible to extract that property from the tens of millions of series contained in the dataset of Google web search queries (Mohebbi et al., 2011).

This is also an example of *crowdledge*, that is, an unexpected knowledge (that the searches for 'wireless hotspot' were more frequent during maxima in solar activity) that emerged from Big Data analysis of individuals' spontaneous digital footprints left in Google searches.

We do not expect strong emergent phenomena to arise from Big Data, as their existence would imply in new fundamental laws of Nature needed to explain them (Chalmers, 2006). As a matter of fact, we are counting on the deductibility aspect of the weak emergent correlations for students to be able to seek scientific explanations for them.

Conclusions

In our understanding, Big Data is more than the largeness of its databases. It was stressed the paramount relevance of the research on the Philosophy of Big Data to deal with its increasing implications in Science, its methods, causality, and Epistemology. Within this philosophy, we emphasised the philosophical concept of 'emergence', both strong and weak ones, and defend our thesis that Big Data Analytics is better understood by the concept of crowdledge, the unexpected knowledge that weakly emerges from Big Data analysis of individuals' digital footprints, than by the overused 5 V definition. The worthwhileness of Science Education is unfortunately contrasted with the current scientific illiteracy and the prevalent school mindset that favours skills and facts against ideas. Inspired on Papert's Constructionism, we presented *learning-with-Big-Data*, a promising way to build scientific knowledge by thinking like a scientist, exploring crowdledge through the mediation of accessible Big Data application software, making concrete abstract concepts such as 'physical laws' and 'causations', what reinforces the need for more conceptually-robust approaches to Dig Data.

Acknowledgements

The author would like to thank Prof. Rodrigo Dalla Vecchia, from ULBRA-Lutheran University of Brazil, and Dr. Melanie Swan, from Kingston University, for many enlightening discussions and suggestions during the development of these ideas and the preparation of this paper.

References

Anderson, C. (2008). The End of theory: The data deluge makes the scientific method obsolete. *Wired – Science*, (16.07). Retrieved 26 November 26 2015, from http://www.wired.com/2008/06/pb-theory.

Ashton, K. (2009). That "Internet of Things" thing: In the real world, things matter more than ideas. *RFiD Journal*. Retrieved 22 June 2015, from http://www.rfidjournal.com/articles/view?4986.

Baase, S. (1996). *A Gift of Fire: Social, Legal, and Ethical Issues for Computing and the Internet*. Upper Saddle River, NJ: Prentice Hall.

Baram-Tsabari, A., & Segev, E. (2009a). Just Google it! Exploring new web-based tools for identifying public interest in science and pseudoscience. In Y. Eshet-Alkalai, A. Caspi, S. Eden, N. Geri, & Y. Yair (eds.), *Proceedings of the Chais Conference on Instructional Technologies Research 2009: Learning in the Technological Era* (pp. 20–28). Raanana: The Open University of Israel.

Baram-Tsabari, A., & Segev, E. (2009b). Exploring new web-based tools to identify public interest in science. *Public Understanding of Science*, 20(1), 130–143.

Baram-Tsabari, A., & Segev, E. (2015). The half-life of a "teachable moment": The case of Nobel laureates. *Public Understanding of Science*, 24(3), 326–337.

Bauer, H. H. (1994). *Scientific literacy and the myth of the scientific method*. Champaign, IL: University of Illinois Press.

Bedau, M. A. (1997). Weak Emergence. In J. E. Tomberlin (ed.), *Philosophical Perspectives, Mind, Causation and World* (Vol. 11, pp. 375–399). Malden, MA: Blackwell.

Bedau, M. A. (2002). Downward causation and the autonomy of weak emergence. *Principia*, 6(1), 5–50.

Bedau, M. A. (2008). Is weak emergence just in the mind?. *Minds and Machines*, 18(4), 443–459.

Bedau, M. A., & Humphreys, P. (eds.). (2008). *Emergence: Contemporary readings in philosophy and science*. Cambridge, MA: MIT Press.

Beulke, D. (2011). *Big Data Impacts Data Management: The 5 Vs of Big Data [Blog post]. Dave Beulke Blog*. Retrieved 7 May 2013, from http://davebeulke.com/big-data-impacts-data-management-the-five-vs-of-big-data.

Bhaskar, R. (1975). *A Realist Theory of Science*. Leeds: Leeds Books.

Bloom, H. K. (1995). *The Lucifer Principle: A Scientific Expedition into the Forces of History*. New York: The Atlantic Monthly Press.

Bloom, H. K. (2000). *The Global Brain: The Evolution of Mass Mind from the Big Bang to the 21st Century*. New York: Wiley.

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679.

Bülbül, M. Ş. (2009). *Google centered search method in pursuit of trends and definitions in Physics and education*. Retrieved 7 February 2014, from http://www.fizikli.com/piwi/fizikli6.pdf.

Bunge, M. (2003). *Emergence and Convergence: Qualitative Novelty and the Unity of Knowledge*. Toronto: University of Toronto Press.

Bunge, M., & Mahner, M. (1997). *Foundations of biophilosophy*. New York: Springer.

Chalmers, D. J. (2006). Strong and weak emergence. In P. Clayton & P. Davies (eds.), *The Reemergence of Emergence* (pp. 244–256). New York: Oxford University Press.

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A survey. *Mobile Networks and Applications*, 19(2), 171–209.

Crutchfield, J. P. (1994). The calculi of emergence: Computation, dynamics and induction. In K. Kaneko, I. Tsuda, T. Ikegami, H. Flaschka & F. H. Busse (eds.), *Proceedings of the Oji international Seminar on Complex Systems: From Complex Dynamical Systems to Sciences of Artificial Reality* (Vol. 75, pp. 11–54). New York: Elsevier.

Dos Santos, R. P. (2014). Big Data as a mediator in science teaching: A proposal. *Innovation Educator: Courses, Cases & Teaching eJournal*, 2(25), 1-13.

Dos Santos, R. P. (2016). *On the definition of Big Data: A systematic review and meta-analysis* (in preparation).

Dos Santos, R. P. & Lemes, I. L. (2016). *Learning-with-Big Data in Science Education* (in preparation).

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V. R., Tsou, A., Weingart, S., & Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, 66(8), 1523–1545.

Feinstein, N. (2011). Salvaging science literacy. *Science Education*, 95(1), 168–185.

Field, H. (2003). Causation in a Physical World. In M. J. Loux & D. W. Zimmerman (eds.). *Oxford Handbook of Metaphysics* (pp. 435-460). Oxford: Oxford University Press.

Floridi, L. (2012). Big Data and their epistemological challenge. *Philosophy & Technology*, 25(4), 435–437.

Fox, P., & Hendler, J. (2014). The Science of Data Science. *Big Data*, 2(2), 68-70.

Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, 66(4), 651–661.

Gartner Group (2015). *Gartner's Hype Cycles for 2015: Five Megatrends Shift the Computing Landscape*. Stamford, CT: Gartner Group. Retrieved 6 September 2015, from https://www.gartner.com/doc/3111522.

Goldstein, J. (1999). Emergence as a construct: History and issues. *Emergence*, 1(1), 49–72.

Google Inc. (2015). *Google Inc. Announces Second Quarter 2015 Results*. Retrieved 12 December 2015, from http://investor.google.com/earnings/2015/Q2_google_earnings.html.

Gray, J. (2007). *eScience - A Transformed Scientific Method*. Retrieved 12 December 2015, from http://research.microsoft.com/enus/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf.

Hacking, I. (2012). *Introductory essay in Thomas S. Kuhn (ed.) The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.
Healy, L., & Kynigos, C. (2009). Charting the microworld territory over time: Design and construction in mathematics education. *ZDM Mathematics Education*, 42(1), 63–76.
Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.
Higginbotham, S. (2011). *Data for doctors: Big data meets a big business*. Retrieved 15 May 2013, from http://gigaom.com/2011/08/08/data-for-doctors-big-data-meets-a-big-business.
Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big Data for Dummies*. Hoboken, NJ: John Wiley & Sons.
IBM (2011). *What is big data?* Armonk, NY: IBM. Retrieved 10 May 2013, from http://www-01.ibm.com/software/data/bigdata.
Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 1–12.
Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.
Kynigos, C. (2012). Constructionism: Theory of Learning or Theory of Design? In S. J. Cho (ed.), *Proceedings of the 12th International Congress on Mathematical Education* (pp. 417-438). Cheongju: Korea National University of Education.
Landemore, H., & Elster, J. (2012). *Collective Wisdom: Principles and Mechanisms*. New York: CUP - Cambridge University Press.
Laney, D. (2001). *3-D Data Management: Controlling Data Volume, Velocity and Variety (No. 949)*. Stanford, CT: Gartner Group. Retrieved 15 May 2013, from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
Lévy, P. (1997). *Collective Intelligence: Mankind's Emerging World in Cyberspace.* Cambridge, MA: Plenum Trade.
Lewes, G. H. (1875). *Problems of Life and Mind*. Boston and New York: Houghton Mifflin.
Mayer-Schönberger, V., & Cukier, K. (2014). *Learning With Big Data*. Boston, MA: Houghton Mifflin Harcourt.
Mohebbi, M. H., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H., & Kumar, S. (2011). *Google Correlate Whitepaper*. Retrieved 12 May 2015, from http://www.google.com/trends/correlate/whitepaper.pdf.
NOAA/NESDIS/NCEI (2013). *Sunspot Numbers - International*. Retrieved 2 May 2013, from http://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-indices/sunspot-numbers/international/listings.
Norvig, P. (2009). *All we want are the facts, ma'am*. Retrieved 14 September 2015, from http://norvig.com/fact-check.html.
O'Connor, T., & Wong, H. Y. (2015). Emergent Properties. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved 14 September 2015, from http://plato.stanford.edu/entries/properties-emergent.
O'Leary, D. E. (2013). Big Data, the Internet of Things and the Internet of Signs. *Intelligent Systems in Accounting, Finance and Management*, 20(1), 53–65.
Papert, S. A. (1980). *Mindstorms - Children, Computers and Powerful Ideas*. New York: Basic Books.
Papert, S. A. (1987). Information technology and education: Computer criticism vs technocentric thinking. *Educational Researcher*, 16(1), 22–30.
Papert, S. A. (1993). *The Children's Machine: Bringing the Computer Revolution to our Schools*. New York: Basic Books.
Papert, S. A. (1999, March 29). Papert on Piaget. Retrieved 12 December 2015, from http://www.papert.org/articles/Papertonpiaget.html.
Papert, S. A. (2000). What's the big idea? Toward a pedagogy of idea power. *IBM Systems Journal*, 39(3.4), 720–729.
Papert, S. A., & Harel, I. (1991). Situating Constructionism. In I. Harel & S. A. Papert (eds.), *Constructionism: Research Reports and Essays, 1985-1990* (pp. 1–14). Norwood, NJ: Ablex Publishing.
Parmaxi, A., & Zaphiris, P. (2014). The evolvement of constructionism: An overview of the literature. In P. Zaphiris & A. Ioannou (eds.), *Proceedings of the First International Conference, LCT 2014* (pp. 452–461). Cham (ZG), Switzerland: Springer.
Prensky, M. (2009). H. Sapiens Digital: From Digital Immigrants and Digital Natives to Digital Wisdom. *Innovate: Journal of Online Education*, 5(3), 1.
Rathod, P. A., Pamdit, K. N., Khandel, H. R., & Raut, T. V. (2015). Internet of Things. *Journal Data Mining Knowledge Engineering*, 7(9), 297–301.
Reichenbach, H. (1956). *The Direction of Time*. Berkeley, CA: University of Los Angeles Press.
Rosa, M. (2008). *A Construção de Identidades online por meio do Role Playing Game: relações com o ensino e aprendizagem de matemática em um curso à distância*. PhD Thesis, Universidade Estadual Paulista, Rio Claro, SP, Brazil.
Segev, E., & Baram-Tsabari, A. (2012). Seeking science information online: Data mining Google to better understand the roles of the media and the education system. *Public Understanding of Science*, 21(7), 813–829.
Snijders, C., Matzat, U., & Reips, U.-D. (2012). "Big Data": Big gaps of knowledge in the field of Internet Science. *International Journal of Internet Science*, 7(1), 1–5.
Surowiecki, J. (2005). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Doubleday.
Swan, M. (2015a). Philosophy of Big Data: Expanding the Human-Data Relation with Big Data Science Services. In D. Webster (ed.), *Proceedings of the IEEE BigDataService* (pp. 468-477). Washington, DC: IEEE Computer Society.

Swan, M. (2015b). Digital Simondon: The collective individuation of man and machine. *Platform: Journal of Media and Communication*, 6(1), 46–58.
Swan, M. (2015c). We *should consider the future world as one of multi-species intelligence*. Retrieved 10 December 2015, from http://edge.org/response-detail/26070.
Swan, M. (2016). The future of brain-machine interfaces: How to feel comfortable joining a cloudmind collaboration. *The Journal of Evolution and Technology* (In review).
Van Fraassen, B. C. (2004). Science as representation: Flouting the criteria. *Philosophy of Science*, 71(5), 794–804.
Van Fraassen, B. C. (2008). *Scientific Representation: Paradoxes of Perspective. Metaphilosophy*. Oxford: Oxford University Press.
Vintiadis, E. (2013). Emergence. In J. Fieser & B. Dowden (eds.), *Internet Encyclopedia of Philosophy*. Retrieved 13 December 2015, from http://philpapers.org/archive/VINE.pdf.
Vitinskii, Y. I. (1965). *Solar Activity Forecasting (No. NASA TTF-289)*. Washington, DC: NASA.
Wan, P. Y. (2009). Emergence a la Systems Theory: Epistemological Totalausschluss or Ontological Novelty? *Philosophy of the Social Sciences*, 41(2), 178–210.
Wilensky, U. (1991). Abstract meditations on the concrete and concrete implications for mathematics education. In I. Harel Caperton & S. A. Papert (eds.), *Constructionism: Research Reports and Essays, 1985-1990* (pp. 193–204). Norwood, NJ: Ablex.
Wivagg, D., & Allchin, D. (2002). The Dogma of "The" Scientific Method [Guest Editorial]. *The American Biology Teacher*, 64(9), 645–646.
Woodcock, B. A. (2014). "The Scientific Method" as Myth and Ideal. *Science & Education*, 23(10), 2069–2093.
Woolley, A. W., Chabris, C. F., Pentland, A. S., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004), 686–688.
Yin, C., Sung, H.-Y., Hwang, G.-J., Hirokawa, S., Chu, H.-C., Flanagan, B., & Tabata, Y. (2013). Learning by searching: A learning environment that provides searching and analysis facilities for supporting trend analysis activities. *Journal of Educational Technology & Society*, 16(3), 286–300.
Zimmerman, C., & Croker, S. (2014). A prospective cognition analysis of scientific thinking and the implications for teaching and learning science. *Journal of Cognitive Education and Psychology*, 13(2), 245–257.

To cite this article: dos Santos, R. P. (2015). Big Data: Philosophy, emergence, crowdledge, and science education. *Themes in Science and Technology Education*, 8(2), 115-127.

URL: http://earthlab.uoi.gr/theste